International Journal of Educational Science and Research (IJESR)
ISSN(P): 2249-6947; ISSN(E): 2249-8052
Vol. 4, Issue 4, Aug 2014, 29-36
© TJPRC Pvt. Ltd.

ESTABLISHING VALIDITY AND RELIABILITY OF ACHIEVEMENT TEST IN BIOLOGY FOR STD. IX STUDENTS

SURUCHI[1] & SURENDER SINGH RANA[2]

[1]Research Scholar, M. D. University, Rohtak, Haryana, India

[2]Associate Professor, T. R. College of Education, Sonepat, Haryana, India

ABSTRACT

This piece of work is focussed on the establishment of validity and reliability of an achievement test designed for class IX students in Biology subject as per CBSE course. After the try-out of initial draft of test on 500 students and item analysis, 111 test items were finalised for the final draft of the test. The final draft of the test was considered for reliability and validity analysis. Reliability of the test was established by Split-half method using Spear-Brown prophecy formula and Cronbach's alpha coefficient. Reliability of the test was found to be good with alpha coefficient value of 0.94. Face validity and Content Validity of the test was established by a panel of ten subject experts giving ratings to the quality of test on a 5-point Likert scale. Overall rating to the test was also good with an average of 4.5. Current study is significant for new researchers seeking information regarding the setup procedure for reliability and validity. School teachers may also use the achievement test to evaluate knowledge, understanding, application and skill of students in Biology.

KEYWORDS: Validity, Reliability, Cronbach's Alpha, Split-Half Reliability, Spear-Brown Prophecy Formula

INTRODUCTION

An Achievement test is developed by a teacher, researcher or educationists to measure the level of skill, knowledge or understanding about a certain topic in a specific area at a given time. Whether the test is indeed measuring what it is intended to measure and whether repeated measurements of the test produce same results under same novice conditions are the questions which are well responded by constructing standardised tests. Standardised tests show a relative degree of validity and reliability which ensures the consistency and stability of test scores every time the test is conducted. Just because a test is reliable, it is not necessarily valid and vice-versa. Thus in any research work, it is essential to establish both the reliability and validity of measurements being used to get the real impact of results.

Validity

Validity refers to how well a test measures what it is intended to measure. **Samuel Messick[1](1988)** defines validity as an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inference and actions based on test scores and modes of assessment. Validity is a matter of degree, not absolute valid or absolute invalid. Thus, while determining the validity of a test, it is essential to study the test results in the settings in which they are used. In any research study validity has two essential parts.

- **Internal Validity:** It refers to the outcome of results from two- group and multi-group design because of the way groups are selected, data are recorded and analysis is performed. If the groups are not treated likewise internal validity is poor. “Approximate validity with which we infer that a relationship between two variables is causal” (**Cook and Campbell[2], 1979).**
- **External Validity:** According to **Last[3] (2001),** External validity, often called “Generalizability”, involves whether the results given by the study are transferable to other groups of interest. A study conducted for a particular gender, race or geographical subgroups may not be relevant to another. Without internal validity, external validity cannot be achieved.

Crocker and Algina[4] (1986) and Brown[5] (1996, p 231-249) categorised validity into three types-

Content Validity

It is the match between test questions and the content of subject to be measured. For content validity, Face validity and curricular validity should be studied.

- **Face Validity:** It is the extent to which a test is accepted by the teachers, researchers, examinees and test users as being logical on the “face of it”. Though it’s a weak form of establishing validity yet can be improved by making it more systematic.
- **Curricular Validity:** It is the extent to which the content of the test matches the objectives of a specific curriculum. It is evaluated by group of curriculum or content experts to judge the proper balance between test items and curricular objectives.

Criterion-Related Validity

It is about the relationship between a test score and an outcome or criterion. The test scores are correlated to the criterion to determine how well they match the criterion behaviour. It can be either predictive or concurrent validity.

- **Predictive Validity:** It refers to the usefulness of the tests scores to predict future events or performance.
- **Concurrent Validity:** It is measured when the test scores and the criterion measure are either made concurrently or in close proximity to each other. It needs to be examined when one measure is substituted for another. If the correlation between the test scores and the criterion is strong, test is said to be valid.

Construct Validity

It is the degree to which a test is consistent with the underlying theoretical concepts being measured. It can be assessed by a panel of experts familiar with the construct. Though construct validity is never fully established.

- **Convergent Validity:** It demonstrate that two tests of same construct correlate strongly in measuring closely related skills or knowledge. High correlation shows the degree to which one measure is similar (converges) to other measure.
- **Discriminate Validity:** It demonstrate a low correlation between two tests measuring different construct showing the degree to which one measure is not similar(diverges from) other measure of different construct.

Common Threats to Validity

Common threats to internal validity are ambiguous temporal precedence, history, maturation, testing, Instrumentation, regression artefacts or regression to the mean, differential selection of sample and mortality or dropping out of participants. Common threats to external validity are aptitude-treatment interaction, all situational specifies like treatment conditions, time, location, noise, administration etc., pre-test and post- test effects, Hawthrone effects (cause-effect relationship not generalizable to other settings or situations), Rosenthal effects (cause-consequence relationship may not be generalizable to other investigators or researchers).

Ways to Improve Validity

Common ways to improve validity are clearly defined and operational goals and objectives, proper review and feedback by experts, review by students for troublesome wordings and other difficulties faced by them.

Reliability

Gay[6] (1987), Reliability is the degree to which a test consistently measures whatever it measures. **Nunnally[7] (1967)** defined reliability as the extent to which measurements are repeatable and that any random influence which tends to make measurements different from occasion to occasion is a source of measurement error". There are three main aspects of reliability viz. Equivalence, Stability and Internal consistency. Equivalence deals with "parallel forms" and "Inter-rater ability". Stability can be assessed through a "Test-Retest" procedure and internal consistency can be established through "Average Inter-item correlation" and "Split-Half Reliability".

Equivalent Forms or Parallel Forms Reliability

It is the measure of reliability which is obtained by administering two parallel forms of tests both of which contain items determining same construct skills, knowledge etc. Except for the actual items. Test scores from both the tests are correlated to evaluate the consistency of results. Practically, Parallel or equivalent forms are rarely instrumented as it is quite difficult to verify the equivalence of two tests.

Inter-Rater Reliability

It refers to the degree of consistency to which observers or raters or judges agree in their judgements on test scores.

Test-Retest Reliability

It is the degree to which test scores are consistent from one testing session to another testing session. Test scores from all the sessions conducted are correlated to establish the stability of measurement over time.

Internal Consistency Reliability

It refers to the extent to which different test items are measuring the same thing and produce similar results. It has two subtypes

- **Average Inter-Item Correlation:** It is obtained by determining the correlation coefficient for each pair of items on the test and taking the average of all of determined correlation coefficients. If the test items are highly correlated with each other, reliability of the entire measure is high.

- **Split-Half Reliability:** It is obtained by splitting in the entire test items into two sets like odd/ even items or first half of the items and second half of the items. The entire test is administered to same group of individuals and the two total “set” scores are correlated to establish the reliability of the test. If the correlation is high, reliability of the test is high.

Major Threats to Reliability

- Small size of sample
- Too less or too many test items
- Poorly constructed test items
- Unconventional test administration
- Subjective scoring
- Very high or very low difficulty of test items
- Student factors like fatigue, illness, anxiety etc.

Ways to Improve Reliability

- Writing longer tests by adding good quality questions
- Sample size should be large.
- While constructing a test, difficulty and discrimination level of items should be considered.
- Standard administrative procedures should be followed; giving clear directions on the test.

None of the researcher can possibly avoid all the threats to validity and reliability of a test but can minimize these by following study protocol.

OBJECTIVE OF THE STUDY

To establish the validity and reliability of an achievement test constructed by the researcher for std. IX students in Biology subject as per CBSE course.

METHODOLOGY

Participants

For the present study, a total of 500 students were selected randomly, 250 from three government schools and 250 from three private schools located in the rural and urban areas. The findings are based on the test scores of 475 students as the data provided by 25 students was found to be totally or maximally incomplete to be considered valid for the study.

Instrument Used

An Achievement test of 146 objective type items was constructed by the researcher on three chapters of Biology for class IX students, seeking suggestions from subject experts. While preparing the test due weightage was given to the instructional objectives, subject content and forms of questions. After the second try-out on the selected sample of

500 students, difficulty value and discrimination power of each and every item was found out statistically. After the first and second try-out and statistical item analysis a total of 35 test items were not found in the acceptable criterion thus number of questions were cut down to 111. Eventually, final draft of the test comprised of 111 objective type test items which was considered for establishing reliability and validity.

Data Collection

Second try-out of the test was conducted on a sample of 500 students. Students were given clear instructions and allotted as much time as they required to complete the test. In addition, students were directed time to time regarding their doubts and were observed throughout the test. At the end, answer sheets from students were collected systematically.

Data Analysis

After the assessment of answer sheets it was found that 25 sheets were either blank or too incomplete to be considered as data. Test scores of remaining 475 answer sheets were arranged in descending order using M. S. Excel. Test scores were than divided into top 27% and lower 27% to carry out item analysis 78% of the items fall in the range of 0.51 to 0.80 of difficulty value while 81% of test items had a discrimination power index of 0.40 and above and thus fall in the category of best items. Test items of the final draft were used to carry out the statistical analysis for evaluating reliability and validity of the test.

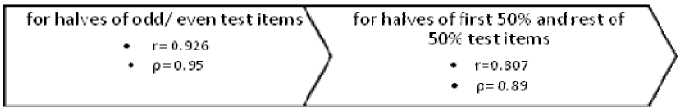
Reliability Analysis

Establishing reliability is the first step towards validation process as a test which is not reliable cannot be a valid one. As the constructed test consisted of only objective type items that are scored as either correct or incorrect and was administered to all the students only once, test reliability was estimated by *measures of internal consistency*. Internal consistency refers to the degree of interrelatedness among the items (**Crano& Brewer[8], 1973; Green et al.[9], 1977)**. Internal consistency can be determined using split-half reliability estimation, coefficient alpha (cronbach alpha) index or Kuder-Richardson formula 20(KR-20) index. Coefficient alpha and KR-20 both represents the average of all possible split-half reliability estimates. KR-20 is used only for dichotomous responses like yes/no or true/false whereas alpha applies to any set of items irrespective of response scale. To establish the reliability of the test, researcher opted for split-half reliability method. The *split-half reliability* of the test was calculated in following steps.

- The test was first divided into two halves. First half consisted of only even test items and second half consisted of only odd test items.
- Pearson correlation coefficient(r) was computed between the scores on the two halves of the test.
- Split-half reliability as the correlation between two halves of the test was calculated using the **Spearman-Brown prophecy** formula, as follows:

$$\rho = \frac{2r}{1+r}$$

Where “r” is Pearson product moment correlation coefficient.



for halves of odd/ even test items

- $r = 0.926$
- $\rho = 0.95$

for halves of first 50% and rest of 50% test items

- $r = 0.807$
- $\rho = 0.89$

For the concerned test computed ‘r’ came out to be 0.9226. After placing the value of ‘r’ in the above formula, reliability of the full test was calculated which came out to be 0.95. This result showed that the test was quite reliable. The main disadvantage of the split-half method is that the two split halves may or may not be equivalent. If they are not, then method underestimates the reliability of the test. This problem was overcome by computing the Spearman-Brown corrected split-half reliability coefficient for other possible split-half and then found the mean of those coefficient. Investigator again divided the test into halves. First half of first 50% questions and second half of other 50% questions for these two halves “r” was 0.807 and “ρ” was 0.89. Cronbach’s alpha is equivalent to average of all estimated split-half reliability coefficients. Formal proof of the equivalence of these two versions of reliability can be found in various text books like **Allen and Yen[10], 1979; Lord and Novick[11], 1968 and Cronbach’s[12], 1951** original article. So, for the test Cronbach’s alpha came out to be 0.92 (α = 0.92). Though the formula for computing the Cronbach’s alpha is:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{sum\ of\ item\ variance}{total\ scale\ variance}\right]$$

Where “k” is number of items.

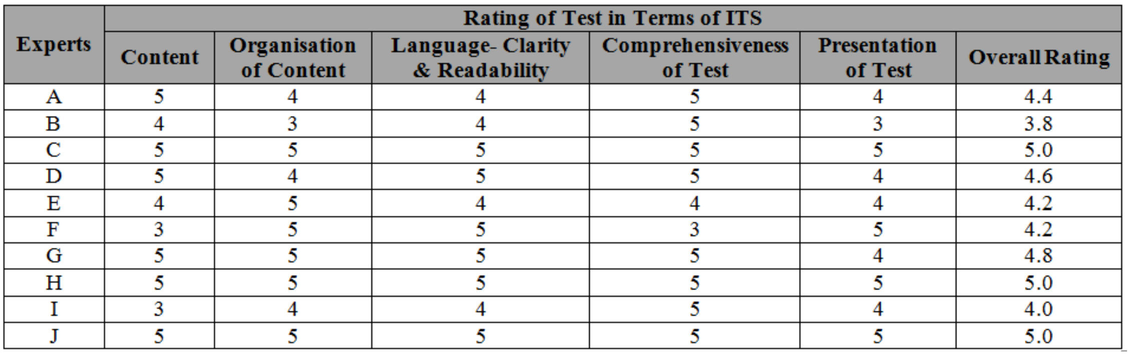
For the test under consideration “k” was 111 and sum of item variance was 24.02 and total scale variance was 383.22. After placing these values in the formula, Cronbach’s alpha coefficient measured to be 0.94. Cronbach’s alpha ranges from 0 to 1.00. One should strive for reliability values of .70 or higher (**Nunnally and Bernstein[13], 1994)**. A commonly accepted rule of thumb is that an alpha of 0.7 indicates acceptable reliability and 0.8 or higher indicates good reliability. Very high reliability (0.95 or higher) is not necessarily desirable, as it indicates a homogenous test with entirely redundant test items. Considering this, the test was found to have quite a good reliability of 0.94.

Test-retest method of reliability was not carried out by the researcher because of the problems of memory, maturation and learning of the content by students during the time gap between two measurements. Equivalent or parallel forms reliability was also ruled out because of the unavailability of an alternate or equivalent form of the constructed test.

Validity Analysis

Assessment of validity cannot be reduced to any one simple technical procedure. Validity is also not something that can be evaluated on an absolute scale. Content validity is usually one of the first ways to ensure the validity of a test or questionnaire or any other measurement. Face validity and Content validity of the test was established by the researcher and a panel of ten subject experts. Test under consideration was given to subject experts for analysis and were requested to rate the quality of the test on Likert scale in terms of the content, organisation of the content, clarity, readability, comprehensiveness and presentation of the test.

Table 1



Experts	Rating of Test in Terms of ITS					
	Content	Organisation of Content	Language- Clarity & Readability	Comprehensiveness of Test	Presentation of Test	Overall Rating
A	5	4	4	5	4	4.4
B	4	3	4	5	3	3.8
C	5	5	5	5	5	5.0
D	5	4	5	5	4	4.6
E	4	5	4	4	4	4.2
F	3	5	5	3	5	4.2
G	5	5	5	5	4	4.8
H	5	5	5	5	5	5.0
I	3	4	4	5	4	4.0
J	5	5	5	5	5	5.0

Ratings on Likert scale for Quality of test

- Extremely poor
- Below Average
- Average
- Above Average
- Excellent

Average of overall ratings from all the ten subject experts came out to be 4.5 showing a good content validity of the achievement test. There is no single best way to study construct validity. To measure the criterion related validity of a test, researcher must graduate it against a well -established standard which acts as a criterion to assess the validity of the new constructed test. No accessibility of such established standard test ruled out the possibility of measuring the criterion related validity of the test.

RESULTS AND DISCUSSIONS

In theory, no study can be banked upon unless there is any mean to conform the standard of data. To improve the quality of any research work it is essential to use a standardised measure or establish the reliability and validity of the used measure or test. Newly constructed Achievement test by the researcher was assessed for its reliability using split-half method and Cronbach's alpha coefficient and found to be highly reliable with the value of alpha 0.94. The test was also found to be valid by a panel of ten subject experts. If the measure one is using is not reliable and valid, their findings cannot be reliable and valid too. Thus, the test designed and developed by the researcher was standardised to be used for further research work.

CONCLUSIONS

This study deals with the procedure to standardise an achievement test, meant for class nine students in Biology subject as per CBSE course, by establishing its reliability and validity after try-out and item analysis. Thus, it can be significant making the teachers aware of the potential use of this achievement test in evaluating the criteria like concept knowledge, understanding, application and skill development in Biology. The study can also be proved useful for the fresh researchers to set up the validity and reliability of their new measures.

REFERENCES

1. Samuel J. Messick (1988). "*The once and future issues of validity: Assessing the meaning and consequences of measurement*", in H. Wainer and H.I. Braun (Eds.), test validity (Erlbaum, Hillsdale, NJ), pp. 33-45.

2. Cook, T. D, & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Boston: Houghton Mifflin.

3. Last, J. (Ed.). (2001). *International epidemiological association. A dictionary of epidemiology (*4th Ed.) New York: Oxford University Press.

4. Crocker, L., and Algina, J. (1986) *Introduction to Classical and Modern Test Theory,* Harcourt Brace Jovanovich College Publishers: Philadelphia

5. Brown, W. (1910), some experimental results in the correlation of mental abilities. *British journal of psychology*, 3, 296-322.

6. Gay, L, 1987. Educational research: *competencies for analysis and application.* Merrill Pub. Co, Columbus.

7. Nunnally, J. C. (1967). *Psychometric theory* (1st Ed.). New York: McGraw- Hill.

8. Crano, W. D, & Brewer, M. B. (1973). *Principles of research in social psychology*. New York: McGraw- Hill.

9. Green, S. B., Lissitz, R. W, & Muliak, S. A. (1977). Limitation of coefficient alpha as an index of test unidimensionality. *Educational and psychological measeurment*, 37, 827-838.

10. Allen, M. J, & Yen, W. M. (1979). Introduction to measurement theory. Monterrey, C. A: Brooks /Cole.

11. Lord, F. M, & Novick, M. R, (1968). *Statistical theories of mental test scores*. Readin, MA: Addison-Wesley.

12. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

13. Nunnaly, J. C, & Bernstien, I. H. (1994). *Psychometric theory* (3rd Ed.). New York: McGraw- Hill.

14. http://research.collegeboard.org/services/aces/validity/handbook/test-validity

15. http://www.uni.edu/chfasoa/reliabilityandvalidity.htm